Creating and exploiting multimodal annotated corpora: the ToMA project

Philippe Blache, Roxane Bertrand, Gaëlle Ferré

Creating and exploiting multimodal annotated corpora: the ToMA project

Philippe Blache(1), Roxane Bertrand(1) & Gaëlle Ferré(2)

(1) Laboratoire Parole & Langage, CNRS & Aix-Marseille Universités
29, av. Robert Schuman. 13100 Aix en Provence
(2) LLING Université de Nantes
Chemin de la Censive du Tertre. BP 81227. 44312 Nantes cedex 3
e-mail: {blache; roxane.bertrand}@lpl-aix.fr, gaelle.ferre@univ-nantes.fr

Abstract. The paper presents a project aiming at collecting, annotating and exploiting a dialogue corpus from a multimodal perspective. The goal of the project is the description of the different parameters involved in a natural interaction process. Describing such complex mechanism requires corpora annotated in different domains. This paper first presents the corpus and the scheme used in order to annotate the different domains that have to be taken into consideration, namely phonetics, morphology, syntax, prosody, discourse and gestures. Several examples illustrating the interest of such a resource are then proposed.

1. Introduction

In recent years, linguists have become aware that a theory of communication describing real interactions should involve the different domains fo verbal and non-verbal description. This is the reason why linguistics and natural language processing have turned to multimodal data where human communication is represented in its entire complexity. By multimodal data, we mean a complex annotation in the auditory-visual domains, not only visual information, Each domain is itself composed of a set of parameters and must be related to the other dimensions of speech. However, annotating such inputs remains problematic both for theoretical and technical reasons. First, we still need a linguistic theory taking into account all the different aspects of multimodality, explaining in particular how the different linguistic domains interact. At the same time, we need to specify a standardized way of representing multimodal information in order to give access to large multimodal corpora, as richly annotated as possible. What is meant by large corpora is however quite a relative notion since in some linguistic fields such as syntax for instance, corpora of several million words are used whereas in prosody where most of the annotations are made manually, a few hours of speech are considered as a large corpus.

This paper describes the first results of the ToMA project[1] which aims at answering these different issues. In this project we propose to specify the different requisites and needs in the perspective of multimodal annotation. Different from many other projects, ToMA does not focus on a specific problem such as information structure, gesture studies or prosody-syntax interaction. Our goal is the development of generic and reusable annotated resources, providing precise annotations in all possible domains, from prosody to gesture. We propose transcription conventions and information encoding as well as tools helping in the annotation process and access to information.

In the first section, we specify a coding scheme adapted for multimodal transcription and annotations. In the second part, we describe the automation of the production of multimodal resources by means of a platform integrating different annotation tools. This platform consists of a sequence of tools leading from raw data to enriched annotations in each linguistic domain. We illustrate the application of this environment by the description of a large multimodal annotated corpus for French. Finally, we present some first results obtained thanks to this resource.

2. Multimodal resources and coding schemes

Several projects address the question of multimodal resources and their annotation. For example, the LUNA project (cf. [Rodriguez07]) focuses on spoken language understanding. The corpus is made of human-machine and human-human dialogues. It proposes, on top of the transcription, different levels of annotation, from morpho-syntax to semantics and discourse analysis. Annotations have been done by means of different tools producing different formats that become interoperable thanks to the use of an interchange format called PAULA (cf. [Dipper05]). SAMMIE (cf. [Kruijff-Korbayova06]) is another project aiming at building multimodal resources in the context of human-machine interaction. Annotations are done using the Nite XML Toolkit (cf. [Carletta03]); they concern syntactic and discourse-level information, plus indication about the specific computer modality used in the experiment. A comparable resource, also acquired following a Wizard-of-Oz technique, has been built by the DIME project (cf. [Pineda02]) for Spanish. In comparison with previous ones, this resource mainly focuses on first-level prosodic information as well as dialog acts.

These three examples are quite typical of multimodal resources development. The main differences with ToMA are first the nature of the source (in our case human-human natural interaction) and second the richness of the annotation (much more exhaustive and precise for ToMA).

Annotating corpora first requires to specify what kind of information it is necessary to represent and how it is organized. This problem consists in defining a coding scheme. Several of them have been developed in different projects such as MATE, NIMM, EMMA, XCES, TUSNELDA, etc. What comes out is that they are very

[1] ToMA stands for "Tools for Multimodal Annotation" (the French acronym is "OTIM"). Project supported by the ANR French agency, involving different partners (LPL, LSIS, LIMSI, LIA, RFC and LLING).

precise in one or two modalities. However, they usually do not cover the entire multimodal domain nor the very fine-grained level of annotation required in every modality. We propose to combine several existing schemes and to extend them so as to obtain a coding scheme that would be as complete as possible.

- *Corpus metadata*: we use a TUSNELDA-like coding scheme ([Tusnelda05]) in which all the information such as speaker name, sex, region, etc. is noted.
- *Morphology and Syntax*: we propose to adapt the Maptask coding scheme for French in the morphological dimension, completed with syntactic relations and properties.
- *Phonetics*: some annotations are a completed version of MATE ([Carletta99]). The phonetic representation is coded in SAMPA.
- *Phonology and Prosody*: we adopt the coding scheme proposed in [DiCristo04] in which prosodic information is annotated both in an automatic (MOmel-Instsint algorithm, [Hirst00]) and manually.
- *Gesture analysis*: we adapt the MUMIN coding scheme ([Allwood05]), which provides an extensive description of gesture forms, but we propose to code gestures and discourse tags separately. The gesture typology is encoded following the scheme proposed in [McNeill05]. A gesture lexicon is compiled from the existing descriptions found in the literature ([Kendon04], [Kipp04], [Krenn04]).
- *Discourse and conversation analysis*: we use the Maptask ([Isard01]) and DAMSL coding schemes, extended to other discourse types such as narration, description, etc. Using the framework of conversation analysis, we also annotate conversational units (turn-constructional units, [Selting00]). We follow the MUMIN coding scheme again to annotate backchannels phenomena.

On top of these schemes, we also take into account different proposals in our encodinglike the one elaborated in Potsdam (cf. [Dipper07]) which covers many of the annotation domains used in ToMA. The following descriptions illustrate, in the manner of the TEI formalization, some annotation conventions at different levels:

Morphosyntax

```
Token::      attributes: orthography
             content: Lex*
```

```
Lex:: attributes: id category lemma rank prob. freq. phon. reference
      content: msd

      category: {Adj Det Noun Pron Adv Prep Aux Verb Conjunction
Interjection Ignored Punctuation Particle Filled pause}
```

Gestures

```
Head::
      attributes:Movement_Type       Frequency       Horizontal_Plane
Vertical_Plane Side_Type
      Movement_Type: {Nod, Jerk , Tilt , Turn , Shake , Waggle ,
Other}
      Frequency: {Single , Repeated }
      Horizontal_Plane: {Forwards , Backwards , Sideways}
      Vertical_Plane: {Up, Down}
      Sid_Type: {Left , Right}
```

Our coding scheme, still under development, proposes then a general synthesis taking into account all the different levels of annotation for multimodal corpora such as phonetics, prosody, syntax, or gestures, as well as annotations at the discourse level (humor, backchannels, narrative units, conversational turns, etc.).

A general coding scheme is of deep importance not only in terms of standardization and knowledge representation, but also for practical matters: it constitutes the basis for a *pivot language*, making it possible for the different tools (Praat, Anvil, etc.) to exchange formats. This is one of the goals of the PAULA format (cf. [Dipper05]). From the same perspective, starting from an adaptation of this format, we are developing tools implementing such interoperability, relying on a translation between the source format of the tool and this language.

3. The ToMA annotation process

Until now, corpus annotation has been essentially based on written corpora, the annotation of oral corpora being very limited. Some transcribed oral corpora exist, but they rarely contain precise phonetic and prosodic information on top of transcription. The Switchboard corpus has been recently annotated in such perspective (see [Carletta04]) and constitutes an exception. As for multimodality, only few initiative try to build large broad coverage annotated corpora, including such level of precision in each domain. The AMI project is one of them (see [Carletta06]), even though the annotations does not seem to be at the same precision level in the different domains. Our project aims at building such large resource, trying to answer to the needs of researches in each domain (in other words being as precise as possible in the annotations) and at the same time making possible the analysis of domain interaction (annotating as many domains as possible). The problem first comes from the lack of annotation tools and second, the difficulty in integrating annotations into a common format.

The ToMA project's aim is to develop a platform providing help at each step of the process, from raw data to high-level annotations. ToMA specifies conventions for manual annotation steps and is based on freely available tools from different sources for the automatic ones. Most of the tools have been developed by the project partners and are adapted for the specific needs of spoken language processing. They will be distributed under the auspices of ToMA, as well as the annotated corpora. The experiment described in this paper has been used for the annotation of a corpus (*Corpus of Interactional Data – CID*) which is already freely available from the CRDO[2]. Figure 1 describes the state of the general process in which the status of each step, automatic (auto) or manual (manual) is specified. We briefly sketch in what follows the main steps of the process:

- *Segmentation in Interpausal-Units*: Transcriptions are made starting from an automatic pre-segmentation of the speech signal into interpausal-units (IPU) that are blocks of speech bounded by silent pauses of at least 200 ms. IPU segmentation

[2] http://www.crdo.fr/

makes transcription, phonetization and alignment with the signal easier. Moreover, speech overlap phases are extracted from IPUs.

- *Transcription*: conventions are derived from [Blanche-Benveniste87] on top of which other information is added (such as elisions, particular phonetic realizations, etc.). From this initial *enriched orthographic transcription* (EOT), two transcriptions are derived: one is phonological, the other is phonetic. The following example illustrates this step:
  - EOT: `et c(e) qu(i) était encore plus le choc c'est que en [fait, faiteu]`
    (what was even a greater shock was that…)
  - Phonologic version: `et ce qui était encore plus le choc c'est que en fait`
  - Pseudo-phonetic version: `et c' qu était encore plus le choc c'est que en faiteu`
- *Phonetization*: This step produces a list of phonemes. After a tokenization, the symbolic phonetizer (see [Di**Cristo**01]) provides a list of tokens and their phonetization labeled in SAMPA. The EOT may sometimes be difficult to use, and a direct phonetic transcription can be, in some cases, simpler for the transcriber; the phonetizer therefore accepts mixed orthographic and SAMPA symbols as an input.
- *Alignment*: The aligner (cf. [Brun04]) takes as input the list of phonemes and the audio signal. It then localizes each phoneme in the signal.
- *Prosody*: Prosodic annotations essentially encode the prosodic categories (intonational and accentual phrases [Jun02]) and the intonation contours associated to them. Such annotations are exclusively done by experts. The intonative level is also encoded with the Momel-Intsint algorithm ([Hirst00]) in an automatic way: from a phonetic representation of the fundamental frequency curve, INTSINT provides a level of surface phonological representation where the melody is represented by a sequence of discrete symbols ([Hirst05]). Because of the lack of consensus on the phonological system in French, we use the MOMEL-INTSINT system which precisely does not suppose any a priori knowledge of the phonological system of the language. The interest to have both manual annotations and automatic INTSINT annotations is to improve INTSINT itself, but also the knowledge, which is still very fragmentary, of the prosodic domains in French.
- *Morphosyntax*: Morphosyntactic annotation is done automatically, using a POS–tagger (LPLsuite, cf. [VanRullen05]) which has been adapted to spoken language. The system has been trained with appropriate data, and custom correcting code heuristics has been developed. It is then checked and corrected manually.
- *Syntax*: We have developed an original statistical parser, adapted to the treatment of spoken data. This is done in two different phases. The first consists in parsing a spoken language corpus by means of a symbolic parser (cf. [Blache05]). In a second stage, the output is corrected manually, the result being a treebank for spoken language. Finally, the statistical parser is trained on these data. The tool we obtain is used in order to generate the trees of the corpora to be annotated automatically. This output also has to be checked manually.
- *Gesture*: The annotation of the gestures made by the participants is being done manually using ANVIL as shown in Figure 3 below. Facial movements (eyebrow,

head), gaze direction and facial expressions (smiles, etc) are encoded as well as hand gestures. For the latter, McNeill's typology [McNeill05] was used (metaphorics, iconics, deictics, beats) and completed with emblems and adaptors. It has also been decided to annotate gesture phases (preparation, stroke, hold, retraction), as well as gesture apex as proposed by [Loehr04], although this annotation will come in a second step.

- *Discourse*: Discourse events (narrative units, speech particles, etc.) are annotated manually in distinct tiers either in Praat or in Anvil depending on the need for video information (for instance, verbal and vocal backchannels were annotated in Praat whereas gestural ones were annotated in Anvil. After all the annotations are made, they were grouped into a single file for the queries to be made). Annotations are created from the aligned orthographic transcription.

4. The CID: a first multimodal annotated corpus in French

The Corpus of Interactional Data is an audio-video recording of spontaneous spoken French (8 hours, 8 pairs of speakers). It features data recorded in an anechoic room and containing 110.000 words. Each speaker of the dialogue is equipped with a headset microphone enabling the recording of the two speakers' voices on two different sound tracks. This enables the study of speech at a phonemic and prosodic level. It also enables the study of overlapping speech which is frequent in spontaneous interactions but seldom analyzed because of the difficulty to separate the intertwined voices of the two speakers a posteriori. Yet, overlapping speech plays a major role in conversation and requires experimental investigations.

In this corpus, we aimed at collecting useful data for the investigation of all the levels under study: the audio quality is optimum and they have been videotaped with a high quality digital camera. The corpus, described in [Bertrand08], has been annotated following the different steps described above.

We then aligned the orthographic and phonetic transcription with the signal and added information from different linguistic fields (prosodic units, intonative contours, morphosyntactic categories, syntactic phrases, etc.). These annotations have been done separately on the audio files and constitute the basis for our present project which consists in the annotation and processing of the corpus from a multimodal / multi-dimensional perspective.

The annotation of the gestures made by the participants is being done manually using ANVIL as shown in Figure 2. The tiers bear different information, from the f0 curve and pitch targets coded with INTSINT (tier 3) to the conversational units (tier 5). Tier 4 encodes the function of the rising contour (RMC: major continuation rise). The following tiers encode the gestural information and the last tier refers to the interlocutor which produces a gestural backchannel signal (nod).

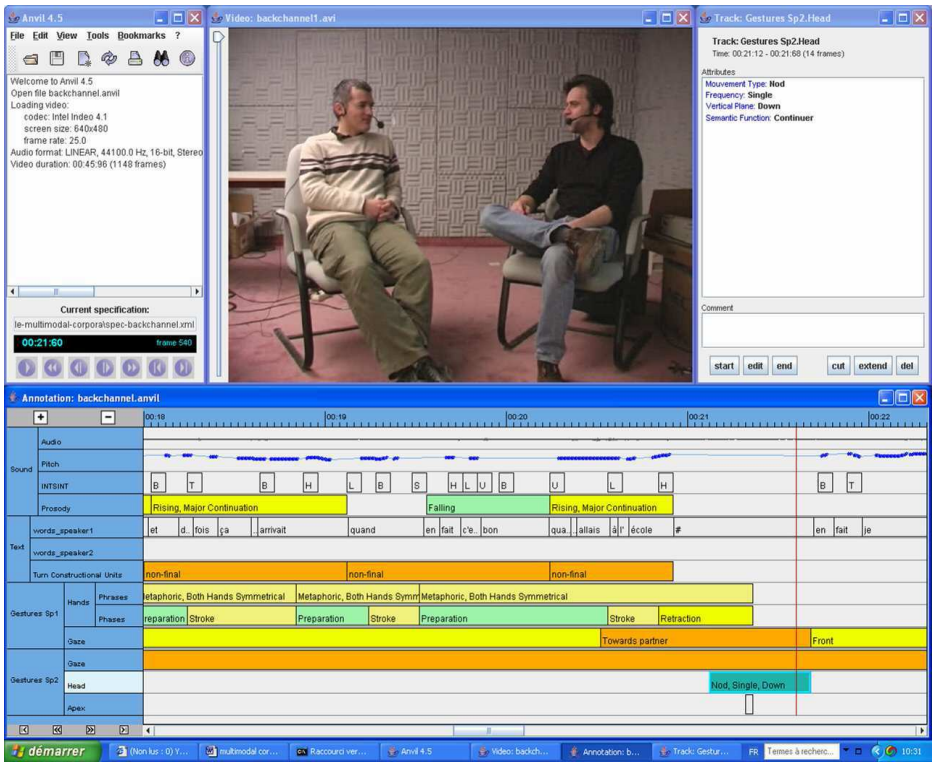


Figure 2: *CID annotation board*

5. Exploiting multimodal annotations

In this section we present several examples to illustrate the kind of information and results that can be obtained thanks to such multimodal resources. In the first subsection we propose some observations which can be done from these data concerning the relations between different levels, in particular prosody and syntax. After this subsection, three studies will be presented: the first study on backchannelling has led to the analysis of some general extenders in French and to the question of reinforcing gestures.

5.1. Data synchronisation

Before entering into data description, it is necessary to tell how data alignment (or synchronization) is done. In our approach, all information is aligned with the signal. This means that identifying interaction between objects requires a temporal comparison. In most cases, an intersection of the temporal segments of the objects to

be considered is the sign of an interaction. For example, an intersection between a deictic gesture and a pronoun specifies a referential relation.

Of course, we have to take into consideration that objects have been aligned on the signal by means of different techniques. In some cases, the alignment is done automatically (e.g. syntactic units aligned with words and phonemes, then with the signal). All the annotations aligned automatically from the phonemes can be strictly aligned in the sense that they share exactly the same boundaries. For example, we have developed a tool segmenting the input on the basis of morpho-syntacitc information. Segments correspond to what we call "pseudo-sentences" that show a certain syntactic coherence. Pseudo-sentences' right boundaries are strictly aligned with that of syntactic groups.

The situation is different when at least one of the annotations has been created manually. In this case, boundaries can be less precise and some flexibility has to be taken into account when identifying the synchronisation between the different annotations. For example, contours and syntactic units usually do not have the same boundaries in our annotations, so that when looking for synchronisation this parameter has to be taken into account.

5.2. Prosody/syntax interaction

Different kinds of prosodic information are available in the CID. At a first level, relations between prosodic units and contours can be underlined. Here are the different categories that have been used:

- *Units* : accentual phrase (`AP`), intonational phrase (`IP`),
- *Contours*: `mr` (minor rising), `m0` (other minor contours), `RMC` (major continuation), `RL` (list rising), `fl` (flat), `F` (falling), `R` (rising), `RF1` (rising-falling), `RF2` (falling from the penultimate), `RQ` (interrogative rising), `RT` (terminal rising).

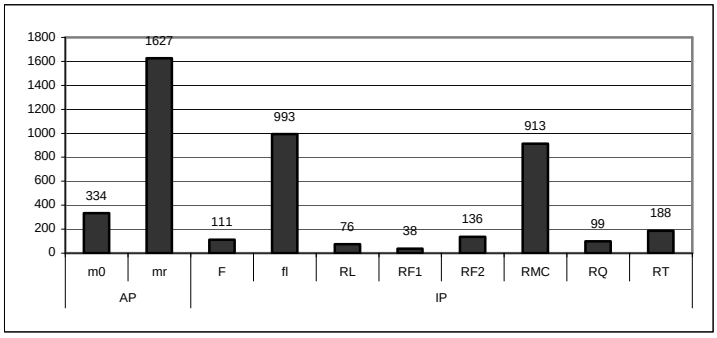
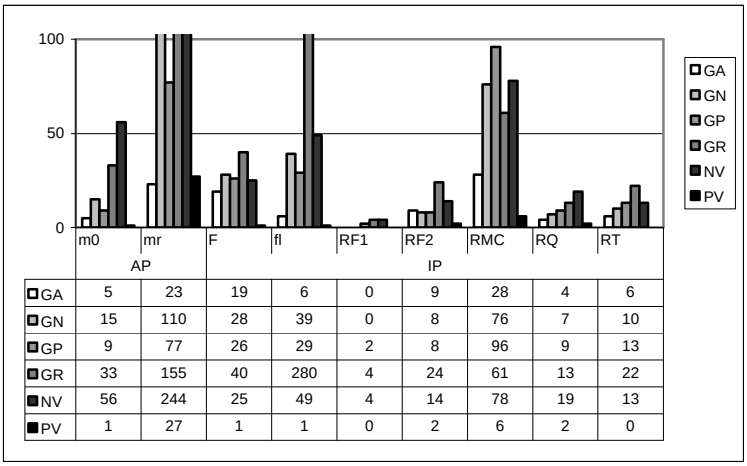


Figure 3: *Relations between contours and prosodic units*

Figure 3 illustrates the distribution of the different contours in one of the dialogue of the CID. This distribution shows that flat and major rising contours are the most frequently used ones at the right boundary of an IP. Conversely, minor rising contours are by far the most frequent type in association with APs.

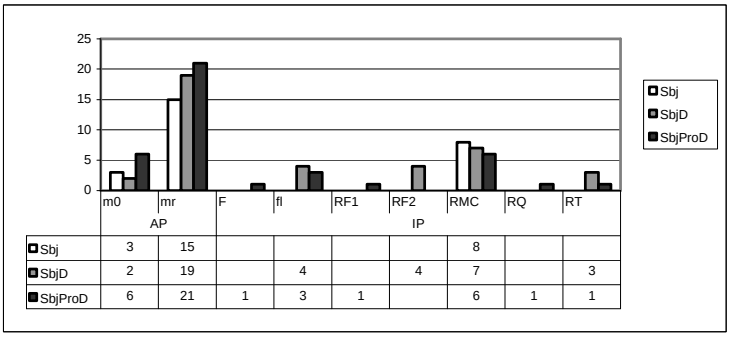


	m0	mr	F	fl	RF1	RF2	RMC	RQ	RT
	AP		IP						
GA	5	23	19	6	0	9	28	4	6
GN	15	110	28	39	0	8	76	7	10
GP	9	77	26	29	2	8	96	9	13
GR	33	155	40	280	4	24	61	13	22
NV	56	244	25	49	4	14	78	19	13
PV	1	27	1	1	0	2	6	2	0

Figure 4: *Relations between syntactic and prosodic contours*

Figure 4 includes syntactic information on top of these prosodic relations. More precisely, it shows the distribution of syntactic units in relation with the different contours. Syntactic annotation of our corpus has been done by means of a stochastic chunker adapted from the techniques developed by the LPL (see [Vanrullen06]). Chunks are non recursive units, defined by the PEAS formalism (see [Paroubek06]) used for the parsing evaluation campaign regularly organized for French parsers. Concretely, we have shown that chunks correspond to a certain kind of supertags (see [Blache08]) or, in other words, identify left boundaries together with the nucleus of the corresponding phrases. The evaluation campaign shows good results for our parser as for spoken language parsing (F-score 80%).

The results of the alignment indicate a strong correlation between /mr/ contours and /NV/ (nucleus VP). This effect can be explained by the fact that these chunks do not contain verbal complements. These complements (in particular direct objects) have a strong syntactic (and semantic) relation with verbs, which explains the fact that no strong prosodic boundary or contour occurs in this place. Reciprocally, chunks corresponding to constituents (such as NP or PP) that usually end main phrases show an important proportion of cooccurrences with intonational phrases.



	m0	mr	F	fl	RF1	RF2	RMC	RQ	RT
	AP		IP						
Sbj	3	15					8		
SbjD	2	19		4		4	7		3
SbjProD	6	21	1	3	1		6	1	1

Figure 5: *Relations between detachments and prosodic contours*

Beside general annotations, the CID also contains more specific ones, added for the study of precise phenomena. For example, detachment constructions have been identified and located (this is an ongoing study, lead by Lisa Brunetti). Among different detachment types, figure 5 focuses on lexical and pronominal dislocated subjects. It shows a relative low level of cooccurrence with intonational phrases and a high proportion of minor rising contours. These data tend to illustrate the syntactic cohesion between the dislocated element and the matrix sentence. This figure also gives indications concerning canonical lexicalized subjects (that are relatively rare in spoken language corpora). What is interesting is that these subjects seem to have the same prosodic characteristics as dislocated ones, including the ones which occur in intonational phrases. This observation should be refined, but it seems that it could illustrate the fact that the detachment construction could become marked in spoken language.

5.3. Backchannels

Backchannel signals (BCs) provide information both on the partner's listening and on the speaker's discourse processes: they are used by recipients to express manifest attention to speakers in preserving the relation between the participants by regulating exchanges. They also function as acknowledgement, support or attitude statement, and interactional signals in marking specific points or steps in the elaboration of discourse. Until recently, they were still considered as fortuitous, but recent works showed that they have a real impact on the speaker's discourse (see [FoxTree99]).

Although they can be verbal like “ouais” (*yeah)*, “ok”, etc, vocal (“mh”) or gestural (nods, smiles, etc), most of the studies on BCs only concern one modality. Our aim is to integrate the different nature of BCs in order to draw up a formal and functional typology. Such information helps in automatically labelling BCs, as well as understanding more accurately the communication strategies (see [Allwood05]). Moreover, we also try to have a better understanding of the BC context which can also inform on its function and contribute to the study of the turn-taking system.

The following example, taken from the CID, illustrates the interest of a multimodal study of BCs. Verbal BCs are represented in italics, gestural ones in frames.

A ah ouais nous on est rentré à (...) dix heures dix heures et demi je crois du soir (...)
B nod
A et elle a accouché à six heures je crois (...)
B *ah quand même ouais*
B head tilt / eyebrow raising
A donc c’était ouais c’était quand même assez long quoi (...)
B head tilt

[A] oh yeah we were admitted at 10, 10.30 I think pm
[A] and she had the baby at 6 I think
[B] [oh yeah right?]
[A] so it was yeah it was quite long indeed

Several questions can be raised: in what context do backchannels appear, do verbal and gestural BCs behave similarly, etc. Such problems require the study of the different levels of information. Among them, the prosodic and discourse layers seem to play an important role for backchannels. Figure 6 shows the relations between these prosodic-unit levels, prosodic contours and conversational turns. By conversational turns, we mean the different units of turn (the turn-constructional units) defined as points of completeness from a syntactic, prosodic and pragmatic point of view [Selting00].
A TCU may be labelled as final (complete from the three criteria), as non final (incomplete from the pragmatic point of view for instance) or as cont(inuation) to refer to cases of adds-on or completion of turn (after a final TCU). As for prosody, the figure shows that BCs are realized preferentially after IPs and /fl/ contours. Concerning discourse, they are realized after final turns, in other words when the speaker reaches a certain level of completion in discourse.

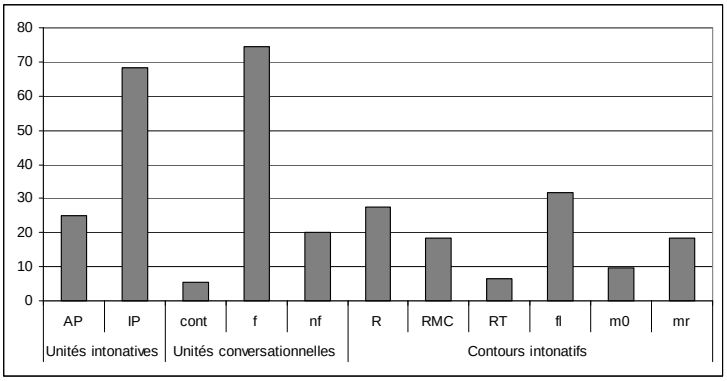


Figure 6: *Relations between backchannels, prosody and discourse*

In [Bertrand07], we have shown that vocal and gestural BCs have similar behavior, in particular concerning the morphological and discursive production context. They appear after nouns, verbs and adverbs, but not after connectors or linking words between two conversational units. As for prosody, gestural BCs can occur after accentual phrases (AP) and intonational phrases (IP) whereas verbal BCs only occur after IPs. Both BCs seem to be favoured by rising and flat contours.
However, rising contours (R, which brings together the whole rising contours RMC + RT) exhibit a specific behavior according to the nature of BC. Proportion tests with a z-score to measure the significant deviation between the two proportions confirm that significant relevant typical contours at points where BC occur are: RT (z-score = 3.23 for vocal BC and 2.18 for gestural BC); RMC (z-score = 2.9 for vocal BC and 4 for gestural BC); and fl (z-score = 2.8 for vocal BC and 3.9 for gestural BC).
By producing preferentially a gestural BC after a RMC contour, the recipient shows that not only does he understand that the speaker has not finished yet but he does not want interrupt him. On the other hand, by producing more frequently a vocal BC after a terminal rising contour, the recipient displays a minimal but sufficient contribution. But thanks to it the recipient also shows his willingness to stay as recipient at a potential transition relevance place. These first results show that different cues at

different levels of analysis are relevant for BCs occurring. More generally they confirm the relevance of a multimodal approach for conversational data and corpus-based investigations.

5.4. Adjunctive general extenders

The study on BCs has led to an analysis of specific French locutions on 3 hours of the corpus. These locutions are a set of *adjunctive general extenders* (cf. [Overstreet00]) such as "(et) tout (ça)" ("*and all that*") and "et cetera", which are favorable contexts for the production of BCs by the hearer. Two issues are at stake concerning them: (1) whether they should or not count as a category of discourse markers (DM), and (2) what their function is. Our aim is to refine existing work mainly based on syntax and discourse analysis, adding prosodic and gestural descriptions of the extenders.

There is yet no consensus concerning the classification of general extenders as DMs (also sometimes called *pragmatic markers* or *pragmatic particles*). Whereas some do consider they are, they do not meet all the criteria defined by [Schiffrin87] and [Fraser99] to enter the category of DMs: for instance, they cannot be inserted at the beginning of an utterance, and their meaning in context is not always different from the core meaning of the locution.

Yet, they fully meet other criteria such as the fact that they cannot stand alone as subject or object of a VP, they show a range of prosodic contours, etc. An intermediate standpoint consists in considering some instances of general extenders as DMs, but not all of them. This is the point of view we adopted in this study, one coder determining the status of DM whenever the locution showed prosodic dependence with the intonation unit which preceded or followed it. This first annotation would of course have to be cross-examined by other coders as well but our preliminary results are quite interesting to mention here.

They show that DMs are almost systematically de-accented (they do not carry nuclear stress and a number of items are phonetically reduced although this is not systematic: for instance "tout ça" [*tu sa*] is often pronounced [tsa] in this context) and usually follow a rising contour, which is not the case of locutions. They are also significantly accompanied with reinforcing gestures, either head movements or hand gestures. As will be shown in the next section, reinforcing gestures reveal discourse structure and this is also the role of "tout ça" in the following example. It is written in capital letters, is accompanied by a metaphoric hand gesture and is produced just after another DM "et tout" with which it forms a single prosodic unit. The example however should not be considered as a case of reduplication of the locution for emphasis.

Example: tu sais tout ce qui était Provence et tout TOUT CA
"you know all the stuff made in Provence and all that TOUT CA"

When it comes to the second issue concerning general extenders, i.e. their linguistic function, we adopted the typology proposed by [Overstreet00] who suggested that the items have three main values:

1. List extenders (extending a list without naming all the items): "ceux qui font les courses ceux qui font la vaisselle et cetera" ("*the people who do the shopping, the ones who wash the dishes et cetera*")
2. Illustration (giving an example): "c'est comme les marrons qu'on bouffe tout ça c'est des châtaignes aussi" ("*it's like the horse chestnuts that we eat and stuff they are indeed chestnuts*")
3. Intersubjectivity (relationships between the participants to the dialogue): "il avait perdu ses parents tout ça" ("*he had lost his parents and stuff*")

Each general extender was assigned a function on the basis of semantics only (on the written script) with Praat by one coder. Out of 104 occurrences of general extenders, only 4 instances could not be decided on. Our aim was to see if the intuitive annotation of the functions of general extenders would meet any pattern in prosody and gesture.

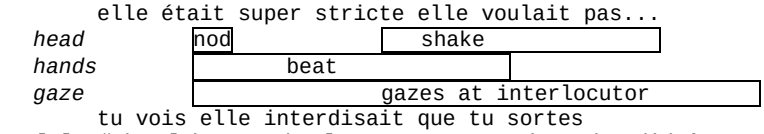
In prosody, we expected to find congruence between the LIST function and the enumerative contour for instance, although the results concerning the correspondence between contours and values need to be developed. But as far as gestures are concerned, we do have preliminary results which are very encouraging. The gestures, which accompany 40 % of the adjunctive general extenders in this corpus, are only head movements (head shakes and head tilts) and hand gestures (metaphorics and iconics).

We never met any eyebrow rising for instance or any smile. We will have to think about such a gestural specificity on general extenders since in other contexts in the corpus, movements and gestures are much more varied. What is more, although head movements were equally distributed among the different functions, we found a much higher proportion of hand gestures reinforcing extenders with an intersubjective value, especially metaphorics.

At last, to loop the loop, since DMs are used to express the pragmatic relationships in dialogue, we expected a higher proportion of DMs than locutions to have an intersubjective value, since this value is the one which is the farthest from the core meaning of general extenders, and this is exactly what our results show.

5.5. Reinforcing gestures

As we have seen in the previous subsection, some gestures can be interpreted as discourse reinforcement devices (cf. [Ferré07], [Ferré09]). To illustrate our point, let us say that there is a difference, for instance, between a head nod produced by the audience as backchannel, and a nod produced by the speaker without any prompt, when this nod doesn't stand for the emblem of "yes". This is the case of the example provided below, where the nod slightly anticipates "super strict", and can be understood as reinforcement of the degree adverb "super".



[A] "she [the teacher] was super strict she didn't want... you see she forbade us to leave the room [during lessons]"

We started with annotating what we intuitively felt were reinforcing gestures, in order to adopt a more gestural perspective rather than a discursive or a prosodic one. Here, we wanted to know what exactly would be reinforced, e.g. instead of focusing on semantic and morphological criteria for intensification, we wanted to find out if there were other possible production contexts for reinforcing gestures. We also wanted to know if gestural and prosodic reinforcement would be simultaneous.

The study showed that intensive gestures are more liable to accompany degree adverbs and negation particles, as well as connectors (DMs which show the discursive or pragmatic links between speech turns). Considering this, we concluded that the gestures we looked at — which were head and eyebrow movements as well as gaze direction — rather played a discursive role of intensification, especially since none of these gestures were associated with any specific stress type. The study also showed that intensive gestures are not redundant in their expression of emphasis: the segments they highlight do not fall under intonational focalization, for instance, with which they are in complementary distribution. This does not mean that reinforcing gestures are never used at the same time as prosodic focalization: in the example above, for instance, there actually is a focal accent on the first syllable of "super" which is also reinforced by the nod. What it means however, is that there are many other contexts, especially negation particles, which are unstressed but yet reinforced by a gesture. Yet, when looking at the gestures themselves and their distribution, one wouldn't speak of emphasis. Indeed, whereas a large gesture could be considered as emphatic (or giving the accompanying speech some emphasis), these movements are not necessarily large at all. Most of the time, they are even very slight. What counts here rather seems to be a question of gesture density, pretty much in the same way as S. Norris [Norris04] speaks of modal density, e.g. the accumulation of body movements on certain parts of speech, listener-oriented.

This study was actually a pilot experiment which must be extended and this will be done in two ways: (a) the rest of the corpus should be annotated in the same way and (b) at the time of the study, the syntactic annotations were not ready to allow their being taken into account, and as they have been done since, they would certainly refine enormously the analysis of the co-occurrences of reinforcement gestures with adverbs and connectors.

6. Conclusion

Annotated multimodal corpora now constitute an essential resource in linguistics. The understanding of language mechanisms (both in production and perception) needs to take into account very precisely the interaction between all the different domains or modalities (phonetics, prosody, lexicon, syntax, pragmatics, gestures,

discourse, etc.). Producing such resources represents however a huge amount of work. It is then necessary to specify a precise framework, identifying the different tasks, the kind of information they have to produce and to what extend they can be automatized.

We have presented in this paper the main characteristics of the ToMA project, providing a general framework for building richly annotated multimodal corpora. The main characteristic of ToMA is that it aims at the description of natural human-human interaction. In order to do this it is necessary to exploit a set of precise and high-level annotations in each linguistic domain. This annotation process has been made possible thanks to a precise definition of different steps, each coming with a set of recommendations and tools.

We have shown in this paper that new results can be obtained from such resources: several examples have been presented here illustrating the importance of a description which brings together information from different levels. It now becomes possible to specify linguistic information in a new perspective, in which phenomena are described in terms of interaction between objects from different domains. This we hope to become an open door for *multimodal grammars*.

References

Allwood J., L. Cerrato, L. Dybkjaer, & al. (2005) The MUMIN Multimodal Coding Scheme, NorFA yearbook 2005, http://www.ling.gu.se/~jens/publications/B%20files/B70.pdf

Bertrand, R., Blache, P., Espesser, R., & al., (2008) « Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle », in revue *Traitement Automatique des Langues*, 49 :3

Bertrand, R., G. Ferré, P. Blache, R. Espesser, and S. Rauzy (2007) "Backchannels revisited from a multimodal perspective", in *Proceedings of Auditory-visual Speech Processing*,

Blache P. & Rauzy S. (2008) «Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks», in proceedings of *TALN08*

Blanche-Benveniste, C. Jeanjean, C. (1987) *Le français parlé, Transcription et édition*, Didier

Brun A., Cerisara C., Fohr D., Illina I., Langlois D., Mella O. et Smaïli K. (2004) « Ants : le système de transcription automatique du Loria», in actes des *XXVe JEP*

Carletta J. & Isard A. (1999) "The MATE Annotation Workbench: User Requirements", in *Proceedings of the ACL Workshop: Towards Standards and Tools for Discourse Tagging*,

Carletta J., S. Evert, U. Heid, J. Kilgour, J. Robertson & H. Voormann (2003) "The NITE XML Toolkit: flexible annotation for multi-modal language data", in *Behavior Research Methods, Instruments, and Computers*, 35:3

Carletta, J. (2006) Announcing the AMI Meeting Corpus. The ELRA Newsletter 11(1)

Carletta, J., S. Dingare, M. Nissim, and T. Nikitina. (2004) «Using the NITE XML Toolkit on the Switchboard Corpus to study syntactic choice: a case study», in proceedings of *LREC04*.

Di Cristo, A. & Di Cristo, P. (2001) « Syntaix, une approche métrique-autosegmentale de la prosodie » in TAL, 42 :1, pp. 69-114.

Di Cristo A., Auran C., Bertrand R., et al., (2004) « Outils prosodiques et analyse du discours », in A.C. Simon, A. Auchlin et A. Grobet (eds), *Cahiers de Linguistique de Louvain* 28, Peeters, pp. 27-84.

Dipper S. (2005) "XML-based stand-off representation and exploitation of multi-level linguistic annotation", in *proceedings of Berliner XML Tage*, Berlin, September 2005.

Dipper S., M. Götze, & S. Skopeteas (eds.) (2007) “Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure” Volume 7 of *Interdisciplinary Studies on In formation Structure*, Working Papers of the SFB 632. University of Potsdam.

Ferré, G, Bertrand, R, Blache, P., Espesser, R. & Rauzy S. (2009) “Gestural Reinforcement of Degree Adverbs and Adjectives in French and English” in proceedings. of *AFLICO*

Ferré, G., R. Bertrand, P. Blache, R. Espesser, and S. Rauzy. (2007) “Intensive Gestures in French and their Multimodal Correlates.” In *Proceedings of Interspeech 2007*

Fraser, B. (1999) “What are discourse markers?” in *Journal of Pragmatics*, 31

Fox Tree, J.E. (1999) "Listening in on Monologues and Dialogues", in *Discourse Processes*, Vol. 27 no. 1,

Hirst, D., Di Cristo, A. & Espesser, R. (2000) *Prosody : Theory and Experiment, chapter Levels of description and levels of representation in the analysis of intonation*, Kluwer

Hirst D. & C. Auran (2005), “Analysis by synthesis of speech prosody: the ProZed environment” in *Proceedings of Interspeech/Eurospeech*

Jun S.-A. & Fougeron C. (2002) «Realizations of accentual phrase in French intonation», *Probus* 14

Kendon A. (2004) *Gesture : Visible Action As Utterance*, Cambridge University Press.

Kipp M. (2004) *Gesture Generation By Imitation. From Human Behavior To Computer Character Animation*. Florida, Boca Raton (http://www.dfki.de/~Kipp/Dissertation.html)

Krenn B. & Pirker H. (2004) “Defining The Gesticon: Language And Gesture Coordination For Interacting Embodied Agents” in *Aisb-2004 Symposium On Language, Speech And Gesture For Expressive Characters*

Kruijff-Korbayova I., C. Gerstenberger, V. Rieser, and J. Schehl. (2006) « The SAMMIE multimodal dialogue corpus meets the NITE XML toolkit” in *proceedings of LREC06*

Loehr D. P. (2004) *Gesture and Intonation*. Doctoral Dissertation, Georgetown University

McNeill D. (2005) *Gesture and Thought*. University of Chicago Press.

Norris, S. (2004) *Analyzing Multimodal Interaction. A Methodological Framework*. Routledge.

Overstreet, M., (1999) *Whales, candlelight, and stuff like that: General extenders in English discourse*, Oxford University Press

Paroubek P., Robba I., Vilnat A. & Ayache C. (2006) «Data Annotations and Measures in EASY the Evaluation Campaign for Parsers in French», *Proceedings of LREC 2006*

Pineda, L. A., Massé, A., Meza, I., Salas, M., Schwarz, E., Uraga, E. and Villaseñor, L. (2002) “The DIME Project”, in *proceedings of MICAI2002*, LNAI 2313

Rodriguez K., S. Dipper, M. Götze, M. Poesio, G. Riccardi, C. Raymond & J. Rabiega-Wisniewska (2007) « Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus”, in *proceedings of Linguistic Annotation Workshop*

Schiffrin, D. (1987) *Discourse Markers*, Cambridge: Cambridge University Press.

Selting M. (2000) “The construction of 'units' in conversational talk”, *Language in Society*, 29,

Tusnelda (2005) *Tübingen collection of reusable, empirical, linguistic data structures.* http://www.sfb441.uni-tuebingen.de/tusnelda-engl.html

Vanrullen T., P. Blache, J-M. Balfourier (2006) “Constraint-Based Parsing as an Efficient Solution: Results from the Parsing Evaluation Campaign EASy”, in *proceedings of LREC06*,